Computational Social Psychology

Fiery Cushman

Department of Psychology, Harvard University, Cambridge, MA, 02138;
email: cushman@fas.harvard.edu

Ann. Rev. of Psych. *in press*. 0:1–30

https://doi.org/10.1146/((to be assigned))

Keywords

formal models; inference; causal attribution; value-guided decision-making; reinforcement learning; game theory; computational social science

Abstract

Social psychologists attempt to explain how we interact by appealing to basic principles of how we think. To make good on this ambition, they are increasingly relying on an interconnected set of formal tools that model inference, attribution, value-guided decision-making, and multi-agent interactions. By reviewing progress in each of these areas and highlighting the connections between them, we can better appreciate the structure of social thought and behavior, while also coming to understand when, why, and how formal tools can be useful for social psychologists.

Contents

1. INTRODUCTION 2
2. INFERENCE 4
   2.1. The Bayesian framework 4
   2.2. Bayesian models of mental state inference 5
   2.3. Bayesian inference beyond theory of mind 7
   2.4. Why Bayes, and why not? 8
3. ATTRIBUTION 8
   3.1. The attribution of causal responsibility 9
   3.2. The attribution of moral responsibility 10
4. VALUE-BASED LEARNING AND DECISION-MAKING 11
   4.1. Reward and value 11
   4.2. Estimating value: Learning and planning 13
   4.3. Value-based Choice 16
5. MULTI-AGENT CHOICE 18
   5.1. Game theory 19
6. CONCLUSION 22

1. INTRODUCTION

Social psychology endeavors to bring two eyes to focus on a common point. One eye is fixed on nature's most beguiling design: Human social thought and behavior. Sociality lays at the heart of human intelligence, cooperation and conflict; it animates politics, history, justice, art and science; it both demands and offers our sense of meaning. Social psychology seeks to explain precisely these things. The other eye is fixed on a particular manner of explanation: One grounded in general principles of human minds. From its inception, social psychology has aspired to explain human's social lives by understanding the workings of each individual's mind—their thoughts, feelings, and actions, and the law-like mental processes that govern them.

As a practical matter, then, it is hardly surprising that a peculiar form of double vision characterizes so many of the field's activities. After all, it isn't easy to discover the general principles of individuals' minds that give rise to the spectacular complexity and power of human sociality. More than one paper, therefore, addresses an expansive topic such as romantic attraction, racial prejudice, or rationalization, by analyzing how people rate the apparent attributes of artificial faces, or by measuring the effect of primes on categorizations, or by recording the neural response to the names of friends and enemies, and so forth. On the one hand, this is exactly what social psychologists must do: to attempt, against all odds, to reconstruct the form of general principles from the fragmentary findings of individual experiments. On the other hand, it often feels as if something is missing: the common focal point where the mode of explanation and object of study meet.

Not just the promise of social psychology, but also its pathologies, were evident from its earliest days. Lewin (1936) lamented the "mere piling up facts" practiced by his contemporaries, which "can only lead to a chaotic and unproductive situation." Heider (1958) echoed Lewin's call: "[W]e shall not attain a conceptual framework by collecting more experimental results. Rather, conceptual clarification is prerequisite for efficient experimentation. Such systematization is an important feature of any science and reveals relationships among

highly diverse events." Like many of their contemporaries, Lewin and Heider thought the way to make progress might be to formalize general principles of individual mental processes in computational terms.

People claim a variety of advantages to computational methods. Sometimes math is useful because it allows us to draw conclusions from premises with quantitative precision. We can determine how long it will take to drive between two cities at a 40 miles per hour; how much weight a bridge will bear; how close a comet will pass to our planet, and on which evening we can glimpse its fiery tail. This works because physical objects like cars, bridges, and comets behave according to sufficiently simple, regular rules. Cognitive operations, however, are so complex and variable, relative to the simplicity of the psychological models used to describe them, that the precise quantitative predictions of those models are rarely taken seriously. This is especially true for the kinds of cognitive operations of interest to social psychologists.

The real advantage of formal models in psychology, then, is not that they afford proof or precision. Rather, as Lewin imagined, it is because they foster conceptual clarity. This is due to three main virtues: Abstraction, convention, and compositionality.

Math is a highly abstract. The same number 2 can characterize a set of apples, ideas, or centuries; the same exponential relation can capture the growth of a bank account or the spread of a virus, and so on. Consequently, math often helps us to see the broad organizing principles of a system in their elemental form.

Of course, many verbal theories in psychology are also highly abstract. In the case of verbal theories, however, this is often considered a disadvantage, because abstraction obscures meaning. For instance, much work supports the idea that cooperation is often our automatic, or intuitive, response (Rand et al. 2014). This is an important insight. But, if it were stated so abstractly and as a verbal theory alone, different peoples' understanding of what counts as "cooperation", "automatic", "intuitive", and so forth, could lead to very different interpretations of the meaning of the theory—and, ultimately, the sorts of protracted and acrimonious scientific debates fueled by mutual misunderstanding.

A second advantage of formal models, then, is that they are built upon conventionally understood concepts and relations: numbers, operators, sets, functions, and so on. By grounding verbal theories in these mathematical conventions, it becomes much easier for people to share a common understanding of their meaning. Thus, for instance, cooperation could be defined as a certain type of behavioral strategy in a certain type of multiplayer game. Although such a definition will be limiting in important ways, it has the virtue of scaffolding mutual understanding. It would also be more clear precisely what kind of evidnece bears on the claim. Math, then, allows abstract ideas to be communicated with less ambiguity about meaning, often bringing us a step closer to Lewin's goal of conceptual clarity.

Finally, these conventions are typically compositional. In other words, math has built a system of concepts and relations that play nicely with each other. In the familiar case of Newtonian mechanics, for instance, the mathematical tools that describe position, velocity, force, and momentum interact sensibly, and can be composed in solving new problems. The same is true in psychology: As we shall see, the computational cognitive models used to describe inference, attribution, value-guided decision-making, and game theory naturally interact with each other, and can be composed to model complex cognitive processes. For the same reason, computational cognitive models have enabled especially fertile exchange with neighboring fields in the social and cognitive sciences: sociology, anthropology, evolutionary

biology, philosophy, linguistics, computer science, economics, and many more.

At their best, computational models can help us to formulate the kinds of abstract theories necessary to ground human sociality in psychological principles, and to do so in way that makes our meaning clear and our progress useful to others. The review that follows is highly selective, focusing on a few areas of outstanding success in order to illustrate the potential of formal models, as well as to reflect on their limitations.

2. INFERENCE

Humans form beliefs that go beyond the immediate sensory data available to us. We hear a squishy thwack and believe that our daughter dropped an egg in the kitchen; we see a fleeting expression and believe our spouse has tired of these dinner guests; we smell a mix of rotten food and wet fur and believe our dog buried a bone in the compost heap. How can we know so much from so little?

Causal model: A model of causal relationships (e.g., "sparks cause fire, smoke does not"), rather than their mere associative or predictive ones (e.g., smoke and sparks both predict fire).

Generative model: A cognitive model that generates data, given background conditions. For instance, given a person's goal as input, it could generate representations of their likely actions.

Abduction: A method of inference in which one favors the hypothesis (i.e., unobserved causes) that best explains the data (i.e., observed effects): "inference to the best explanation"

Several computational frameworks are available, and each of them enjoys considerable support. Perhaps the simplest models depend on statistical association: If the last thwack was a brownie-bound egg gone astray, this one is likely to be, as well (e.g. Rescorla 1972). It is also possible to define a set of deductions, or other rule-like transformations, operating of propositions: "If thwack, then either dropped eggs or water balloons; If kitchen, not water balloon; therefore, egg" (e.g. Newell & Simon 1956). Each of these sorts of approaches plays an important role in explaining certain cases. An especially powerful and productive class of models, however, instead adopts a Bayesian approach (Griffiths et al. 2008).

2.1. The Bayesian framework

As employed in cognitive science, the Bayesian approach is characterized by a few essential features.

1. People represent causal models. That is, they know things like "dropped eggs cause thwacks" (as do thrown eggs, dropped water balloons, etc.), "kids cause dropped eggs" (as do grownups, occasionally), "vacuous chit-chat bores my spouse", and so on.
2. These causal models are probabilistic, and often *generative* , meaning that given certain inputs ("causes") the models generate simulated outputs ("effects") according to their probability of occurrence (Vul & Pashler 2008). In other words, like a plinko board, our causal models allow us to drop mental marbles and see where they land. Generative models don't directly encode statistics like "half the marbles land on the right side". Rather, by dropping lots of marbles, we can estimate the distribution of landing positions.
3. We adopt beliefs about unobserved variables based on how well those beliefs explain the observed ones, given our causal models. This is sometimes called abduction, or "inference to the best explanation". Upon observing new information ("data", d), one updates the probability of each hypothesis h_i about unobserved variables based on the likelihood of the observed data given that hypothesis $P(d|h_i)$, according to one's causal model. This likelihood is balanced against the prior probability of the data $P(h_i)$, as well as the same factors for every other one of the n hypotheses under consideration:

$$P(h_i|d) = \frac{P(d|h_i)P(h_i)}{\sum_{j=1}^{n} P(d|h_j)P(h_j)} \quad 1.$$

Thus, for instance, the data is a "thwack", one hypothesis is a dropped egg in the kitchen, another hypothesis is a dropped water balloon in the kitchen, and a third hypothesis is a dropped loaf of bread. The first and third have high priors $P(h)$, while the second does not—i.e., eggs and bread are often dropped in kitchens, but water balloons only rarely. The first and second have high likelihoods $P(d|h)$, while the second does not—i.e., dropped eggs and water balloons often "thwack", but dropped bread rarely does. On this comparative basis, then, the first hypothesis is assigned the highest probability given the data, $P(h|d)$.

Bayesian models have transformed cognitive science for several reasons. First, they unify an astonishing variety of cognitive acts under a single general framework. The very same abstract model usefully explains perception (e.g., how we infer depth from a flat projection on the retina; Nakayama & Shimojo 1992), memory (e.g., how we reconstruct the details of patchy memories; Gershman 2021), language (e.g., how we infer a speaker's meaning from their words; Goodman & Frank 2016), categorization (e.g., how we infer the taxonomy of animals from observable features; Tenenbaum et al. 2006), theory of mind (e.g., how we infer somebody's goals from their actions; Baker et al. 2009), and much more.

Second, Bayesian models are rational models. That is, they specify how an idealized agent with unbounded cognitive resources would optimally solve inference problems. Of course, humans neither have unbounded cognitive resources, nor are they perfectly designed. Nevertheless, understanding the problem that an organism is trying to solve and the abstract form of an idealized solution often helps to structure inquiry into the actual mechanisms underlying thought and behavior (Marr 1982; Tinbergen 1963).

Finally, as we will see in subsequent sections, the Bayesian formalization of inference is usefully compositional, in the sense that it can be integrated with other formal frameworks for understanding attribution, learning, choice, and so forth. This compositionality owes both to formal nature of Bayesian theories, as well as their substantive commitment that humans reason over probabilistic generative models of causal relations.

2.2. Bayesian models of mental state inference

The application of Bayesian methods to model theory of mind—how people infer mental states, and how they predict thought and behavior given those inferences—is their most significant contribution to social psychology to date (reviewed in Jara-Ettinger 2019).

Decades earlier, however, the seeds of this development were planted by Heider and Simmel (1944) , who vividly demonstrated how limited sensory data automatically prompts people to impute rich mental states—in their case, motives, beliefs, and emotions that one cannot help but read into the movements of geometric shapes acting out a brief morality play in a jerky stop-action film. The goals, beliefs, personalities, and morals of the geometric shapes are not literally present on the screen, but people naturally infer these mental states in order to make sense of their actions.

The process of imputing mental states based on observed actions was first formalized as Bayesian inference in a seminal paper by Baker, Saxe and Tenenbaum (2009). Following much prior work, they proposed that the naïve theory of mind that we use to understand others' thoughts and actions is built around a causal model of a goal-directed rational actor. Thus, one can infer a person's unobserved goal from observations of their actions and their environment:

LEVELS OF ANALYSIS: MARR AND TINBERGEN

Marr (1982) and Tinbergen (1963) each made their most important contributions to the behavioral sciences by helping us to understand what it means to explain a behavior. Both argued that a complete explanation of behavior requires multiple, complementary levels of analysis including both a description of the *function* of the system that produced the behavior as well as its *mechanistic* details.

Marr described three levels of analysis for cognitive systems. At the computational level, one seeks to define the problem a system is designed to solve and the idealized form of its solution: One that is abstract, and often "optimal" or "rational". For instance, at the computational level, the visual system could be considered to solve (among others) the problem of inferring a three-dimensional representation of space given a two-dimensional projection of light on the retina. Bayesian methods abstract the idealized form of this inference problem. As such, they help us to understand the function of the visual system even if they do not describe the particular algorithm that the humans use. Marr distinguished this level of analysis from the "algorithmic" level, which corresponds to the actual computations performed by the cognitive system, and the "implementation" level, which corresponds to the biological mechanisms that instantiate the algorithm.

Tinbergen studied animal behavior through the lens of evolutionary theory. He proposed that a complete description of an animal's behavior required an "ultimate" explanation of the adaptive function of that behavior (i.e., its fitness value and evolutionary history), along with a "proximate" explanation of the specific biological mechanism that produced the behavior and its development across the lifespan. Broadly, Tinbergen's adaptive level of analysis and Marr's computational level share a concern with functional design, while Tinbergen's mechanistic level of analysis and Marr's algorithmic and implementation levels share a concern with the specific causal processes that produce behavior within an individual.

$$P(\text{Goal} \mid \text{Actions, Environment}) \propto P(\text{Actions} \mid \text{Goal, Environment}) P(\text{Goal} \mid \text{Environment}) \quad (2.)$$

In a series of experiments, participants viewed a schematic illustration of a person beginning to walk towards one of several locations, navigating around various obstacles. As the walker's path unfolded, participants were asked to infer their goal. These inferences were elegantly captured by the model's predictions. Subsequent work extended this framework to model inferences about beliefs by setting up an environment in which the depicted people's views of goals could be obscured, and modeling the relationship between perception, belief, and planning (Baker et al. 2017).

This model has been further extended to capture a wide array of phenomena of central interest in social psychology (for review, see FeldmanHall & Shenhav 2019; Jara-Ettinger 2019). For instance, just as Bayesian methods can be used to model how people infer beliefs and desires, they can also be used to model how people infer peoples emotional states (Saxe & Houlihan 2017; Ong et al. 2019), the utilities that people assign to objects and even to each other (Jara-Ettinger et al. 2016; Gates et al. 2021; Liu et al. 2017), and underlying dispositional traits such as friendless, trustworthiness, etc. (Kim et al. 2020; Cushman & Macindoe 2009; Diaconescu et al. 2014, 2017; Shin & Niv 2021). For this reason, Bayesian models of mental state inference are particularly indispensable to computational models of human moral judgment (Jara-Ettinger et al. 2015; Yu et al. 2019; Siegel et al. 2018; Crockett et al. 2021; Kleiman-Weiner et al. 2015, 2017). After all, when making moral judgments,

people are exquisitely sensitive to others' mental states: their intentions, beliefs, motives, traits, etc. (Cushman 2008; Young et al. 2007; Uhlmann et al. 2015). Since these cannot be directly observed, they must be inferred.

Bayesian accounts of mental state inference are also central to many formal models of social learning (reviewed in Gweon 2021). Often we can learn best from somebody by focusing not on the literal form or meaning, but instead by asking what it reveals about them: The message they intended to send, or perhaps the message they intended to conceal. A classic case is figurative speech: Our ability to infer that, when someone steps out into miserable weather and says, "Oh I just *love* a surprise hailstorm", their intended meaning is precisely the opposite (Goodman & Frank 2016). Similarly, Bayesian models have been used to explain how people infer general moral principles from patterns of praise and blame (Nichols 2021; Ho et al. 2017), how people learn concepts from others' examples (Shafto et al. 2014), and how people learn complex tasks from demonstrations (Ho et al. 2021). They have also been used to explain how we can integrate our own evidence with the imperfect advice of potentially fallible others (Vélez & Gweon 2019). Finally, Bayesian models can capture not just how we learn, but also how we teach (Vélez et al. 2023). Effective teachers sometimes use theory of mind recursively, anticipating how learners will use their *own* theory of mind to guess at the teachers' communicative intent (Goodman & Frank 2016; Ho et al. 2019, 2021; Shafto et al. 2014). Thus, for instance, a savvy speaker anticipates when a listener will pick up on her irony; a savvy partner anticipates when "the silent treatment" will be appropriately understood as a sanction; and so on. This recursive embedding of theory of mind ("I know that she'll think that I meant...") is essential to both successful collaboration and strategic competition (Kleiman-Weiner et al. 2016; Shum et al. 2019).

2.3. Bayesian inference beyond theory of mind

Bayesian models have also been used by social psychologists beyond the specific case of theory of mind. For instance, they can be used to adjudicate putative cases of motivated reasoning–i.e., drawing inferences that are not justified by the data because they have some hedonic or practical value. After all, some apparent cases of motivated reasoning may be explained by perfectly rational processes (Kunda 1990). For instance, suppose that you judge a stranger harshly when they act rudely, but judge your best friend leniently given the very same behavior. Does this show a bias in favor of your friend? Perhaps, but another possible explanation is that you simply have a great deal more preexisting evidence favoring the hypothesis that your friend is well-mannered, as compared with the stranger (i.e., different priors). Bayesian models offer a useful tool to formalize and test the various possibilities. By measuring peoples' priors, along with other well-specified decision variables, these methods allow us to quantitatively assess whether reasoning is biased. In cases ranging from trait attribution (Kim et al. 2020; Park et al. 2021) to stereotypes (Cao et al. 2019, 2017) to political beliefs (Tappin et al. 2020), formal models have helped to clearly identify cases that do, and do not, implicate motivated reasoning.

Another foundational construct in social psychology is the group. As a practical matter, however, group boundaries can be hard to define, are constantly shifting, and may not be overt. How, then, do ordinary people discover and define social groups? One potential answer is on the basis of homophily, the tendency of people with similar tastes, interests, values, and beliefs to form friendships and alliances. Bayesian models of structure learning

have been fruitfully applied to this problem, providing an impressive match to experimental evidence on how ordinary people infer group boundaries and membership (Gershman et al. 2017; Lau et al. 2018).

2.4. Why Bayes, and why not?

Domain specific: A representation or computation that is specific to one aspect of thought, or one type of problem. For instance, our "theory of mind" is a domain-specific causal model linking perception, thought, feeling, and action.

It is remarkable that seemingly disparate facets of social thought can be systematically related within a single formal framework. At first blush, it is not apparent that there would be a common computational core to the psychological processes of attributing beliefs and goals, understanding irony, diagnosing motivated reasoning, and discovering the structure of social groups—let alone the structure of visual perception, reconstructive memory, and so on. In fact, these processes are *so* distinct that it raises the question,"Exactly what is shared, and how do we account for their evident differences?" It is a strength of the Bayesian framework—one shared with all formal models—that it can tell us precisely where the similarities and dissimilarities between cognitive mechanisms arise.

Let us begin with the dissimilarities. Bayesian inference depends on a causal model that links latent, unobserved variables to observed data, furnishing the likelihood $p(d \mid h)$. The causal model is domain specific, and there is no reason to expect substantive similarities between, say, the causal model of optics that links objects to their projections on the retina and the causal model of rational actors who pursue ice cream when hungry.

Yet, there is an important abstract relationship between visual perception and mental state inference: In both cases we can use a domain specific causal model in order to infer from data we do observe the most likely explanations in terms of variables that we do not.

Apprehending this shared structure is, curiously, quite helpful when we are forced to confront one of the most distinctive limitations of the Bayesian approach: Its computational cost. Except in special cases, the computations involved in exact Bayesian inference are prohibitively demanding. Much work focuses on finding computationally efficient approximations. As these are discovered, they can be propagated across models of highly diverse cognitive domains due to the recognition of their shared abstract structure. Indeed, because Bayesian methods are used for statistical and AI applications outside of psychology, there is also a fertile exchange of methods across fields (Ruiz-Serra & Harré 2023).

Of course, there are also alternatives to Bayesian inference that allow us to form beliefs that go beyond the immediate data we perceive, often in ways that are especially computationally efficient. For instance, even without inferring others' mental states we can learn from them by simply copying what they do, or by other heuristics (Najar et al. 2020; Wu et al. 2022). And, we can predict peoples' mental states using statistical approaches that do not depend on a causal generative model (Tamir & Thornton 2018; Rabinowitz et al. 2018). A key area for future research is to further specify general formal frameworks for non-Bayesian inference, along with models that explain which kinds of inference procedures are favored in which situations (Ho et al. 2022b).

3. ATTRIBUTION

Inference helps us to establish the facts of a situation. But, even when they agree on facts, people may disagree on matters of responsibility. For instance, suppose that each morning at the bakery Alice is supposed to mix the dough, Betty is supposed to start the ovens, and Carl is supposed to sweep the floor. This morning it snows, and Alice arrives too late to mix

RESOURCE RATIONALITY

Many computational cognitive models reside at Marr's (1982) computational level of analysis, describing an idealized solution to the general problem that a cognitive mechanism is designed to solve. This is the case for most Bayesian models of inference, some models of value-guided decision making, and most game theoretic models of multi-agent decision-making. Often, however, tremendous time and computational effort would be required to implement these idealized models exactly. A well-designed cognitive system will substitute less computationally demanding approximations and heuristics in place of the idealized solutions to various problems (Simon 1956). This idea animates "dual process" models of cognition (Kahneman 2011). Over the past several decades, considerable progress has been made in showing how apparent deviations from idealized models can be understood as sensible exchanges for cognitive efficiency (e.g. Todd & Gigerenzer 2000; Bhui et al. 2021). Increasingly, researchers have begun to formalize the very problem of accuracy/effort trade-offs in computationally precise terms so as to derive the optimal design of cognitive systems operating under resource constraint. This style of model is often termed "resource rational". (Lieder & Griffiths 2020).

the dough. Betty and Carl get their jobs done on time, but the bread is not ready by 9am. Who is responsible? Alice, for being late? The snow, rendering Alice blameless? Carl, for not stepping in to mix the dough after a quick sweep? The answers to these questions are not unknown facts that we can infer, but subjective attributions that we make, given those facts.

Social psychologists have devoted great attention to distinguishing the different kinds of attributions that people make, and to detailing the processes by which we make them (Heider 1958; Kelley 1973; Weiner 1995). Two of the most important attributions are those of causal responsibility and moral responsibility. They govern how we assign responsibility to people versus situations (Ross 1977), how we distribute credit and blame (Weiner 1995), and how we explain complex events to ourselves and others (Lombrozo 2010).

3.1. The attribution of causal responsibility

How do we determine that one event is causally responsible for another? Famously, Michotte (2017) observed that sometimes causal attribution is grounded in perception, such as one when billiard ball strikes another. Perceptual processes cannot, however, account for the kinds of attributions of greatest interest to social psychologists: that the launch of Sputnik caused the space race, heartbreak caused Romeo's suicide, or rehiring Steve Jobs saved Apple from oblivion.

Rather, our best formal models of these sorts of causal judgments depend on counterfactual assessments: Roughly, asking, "If somebody had changed X, would Y still have happened?". Thus, for instance, we conclude that Steve Jobs saved Apple from oblivion because we are able to use a causal generative model to consider what *would* have happened to Apple *without* Steve Jobs, and the answer is "oblivion". Causal attributions for physical systems often depend on counterfactual reasoning (Gerstenberg et al. 2021), but evidence suggests that it is particularly important when reasoning about humans' intentional actions (Lombrozo 2010). These psychological accounts of causal attribution build directly on seminal work in philosophy (Lewis 2013) and mathematics (Pearl 2009).

Counterfactual: A specification of a causal model in which one intervenes on one or more variables and then adjusts causally downstream consequences as necessary.

Counterfactual models of causal attribution also share several features with the Bayesian framework described above: Both posit that people represent and reason over causal generative models, and both provide a natural explanation for the probabilistic format of these reasoning processes. Bayesian inference and counterfactual reasoning are, however, quite distinct computations performed over this common representational substrate.

Although counterfactual judgments are crucial to causal attribution, however, they do not provide a sufficient account by themselves (Hitchcock & Knobe 2009). A moment's reflection reveals that a simple, pure counterfactual test ("would the effect have happened without the cause?") is not a very good model for how we ordinarily talk about causal responsibility. To choose just one illustrative example, suppose that somebody wins a lottery because they were able to (1) name Disney's mouse mascot ("Mickey") and (2) guess a secret number between 1 and 1,000,000 ("355,449"). Both (1) and (2) were necessary for the win (as our causal model informs us) but, intuitively, it feels like (2) is what *really caused* the win.

Kelley's (1973) covariation model was perhaps the first formal treatment of this phenomenon, along with several related quirks of causal attribution. Kelly proposed that people attribute responsibility to events that tend to *covary* with an observed outcome. (For instance, naming Mickey Mouse will not be correlated with winning the lottery described above, but guessing the correct number will). A recent formal model pursuing a broadly similar approach can indeed capture nuanced patterns of responsibility judgments (Quillien & Lucas 2022).

A key issue for future research is to determine how these sorts of statistical models can be integrated with, or explained by, models of causal attribution built around the machinery of couterfactual reasoning. Several such models have been developed in recent years, and with promising results (Icard et al. 2017; Morris et al. 2018; Halpern & Hitchcock 2015). In parallel, there has been a burst of activity exploring the *rational* basis of causal attribution: In other words why, at the "computational" (Marr 1982) or "ultimate" (Tinbergen 1963) level of analysis, do we attribute causation the specific sorts of things we do? Several related proposals center on the idea that attributing causes to past events eventually guides our selection of promising future interventions (Morris et al. 2018; Hitchcock & Knobe 2009; Lombrozo 2010; Vasilyeva et al. 2018).

3.2. The attribution of moral responsibility

In at least two ways, the attribution of moral responsibility is intimately linked to the attribution of causal responsibility. First, moral judgment—especially judgments of blame and punishment—often depend on attributions of causal responsibility for harm (Cushman 2008). In other words, if you are judged to have harmed somebody then, all else being equal, you are likely to receive greater blame and punishment.

The second connection, while less obvious, exerts even more influence. We blame people for the harm they *cause*, but also—and especially—for the harm they *intend*. Now, it might seem quite natural to suppose that when we judge a person to have harmed intentionally, this is just a straightforward matter of mental state inference—i.e., a problem to be solved by the Bayesian methods discussed in the prior section, rather than anything to do with causal attribution. Recall, however, that our intuitive understanding of others' mental states is a domain-specific causal model linking perception, thought, and action. Motivated by this observation, several recent formal models propose that what we mean when we say

that a person has caused an event "intentionally", is that we attribute *causal responsibility* for their action to their *preferences* (Quillien & German 2021; Halpern & Kleiman-Weiner 2018; Kleiman-Weiner et al. 2015). That is, intentionality judgments depend in part on the counterfactual claim, "if this person had different preferences, they wouldn't have done that." In a broadly similar spirit, recent work suggests that our ordinary notion of free will (typically a requirement for full moral responsibility) depends upon domain-specific causal attributions of action to thought (Knobe & Nichols 2011; Martin & Cushman 2016). Thus, the mental states that we use to describe and explain others' behaviors—and, crucially, to morally evaluate it—are a product not just of causal inference, but also of causal attribution.

Over the past fifteen years several theorists have made preliminary attempts to assemble formal models of the attribution of moral responsibility in humans (Halpern & Kleiman-Weiner 2018; Kleiman-Weiner et al. 2015; Mikhail 2007; Sosa et al. 2021), often drawing inspiration from well-specified but ultimately more qualitative models (Malle et al. 2014; Alicke 2000; Weiner 1995). In one way or another, most of these integrate both forms of causal attribution described above: A broad causal connection between a person and a harmful outcome, together with a much more specific causal connection between certain mental states and action. But there is an additional ingredient to these models of moral judgment that we have not yet discussed, and to which we will turn our attention next: Value.

4. VALUE-BASED LEARNING AND DECISION-MAKING

No class of computational model has attracted greater attention from social psychologists than value-guided decision-making (Ruff & Fehr 2014; Rilling & Sanfey 2011). These models describe a particular method of action selection; that is, of determining which actions to perform when. Of course, people have many ways of choosing actions without representing their value. For instance, innate reflexes and instincts sometimes map from stimuli to the appropriate action (e.g., withdrawing from a precipice), but without representing value. We might also copy others' behaviors without representing their value. Or, we might repeat actions we've performed in the past, or actions that have been rewarded, without storing any representation of their expected value. Each of these mechanisms for action selection enjoys empirical support (as do others), and several of them have been elegantly formalized. This review focuses on value-based methods of action selection because they have been the most intensively studied and, to date, the most productive formal models employed by social psychologists.

Reward: A numerical objective assigned to states of affairs, or features of them, that an agent seeks to maximize.

Value: T he long-run expected reward associated with performing some action[1].

4.1. Reward and value

A recurring idea across many fields—economics, computer science, psychology, neuroscience, and more—is that agents can represent objectives on a continuous numerical scale, such as utility, reward, or value. This idea can be formalized in different ways. Here, I use the term "reward" to mean a real number assigned to some state of affairs (e.g., "eating a great salad" = 7) or features of it (e.g., "arugula" = 1; "sharp vinaigrette" = 3; "candied pecans" = 3, etc.). In order to obtain reward an agent can take various actions, and the expected long-run reward of an action (given an expected series of further actions) is its "value". Thus, when engaged in value-based decision-making, the goal of an agent is to estimate the values of the available actions and to then choose the highest.

Put somewhat differently, rewards are intrinsic (a special class of things we want "for themselves") whereas value is instrumental (anything could be valuable, but only insofar as it eventually brings reward). To illustrate this idea in the simple "bandit" setting where the task is to choose just one action, suppose than an agent faces a number of actions $a_1, a_2 \dots a_n$ each of which probabilistically transitions to states $s_1, s_2 \dots s_m$, with reward given by $R(s)$. Then, the value Q_a of each action is an average of the rewards of its potential consequences, weighted by their probabilities of occurring given a:

$$Q_{a_i} = \sum_{j=1}^{m} R(s_j)p(s_j \mid a_i), \tag{3.}$$

and the agent should choose the value-maximizing action $\arg\max_i V(a_i)$. This is the basic intuition behind expected utility theory, an important foundation of much work in economics and psychology. The simplicity with which this normative standard can be expressed, however, belies the complexity of actually computing it—especially in cases of sequential or multi-agent action selection. Much work on value-guided decision-making consists in determining how people estimate the value of actions in a computationally efficient manner.

4.1.1. Social reward and social value. Perhaps because expected utility theory rose to prominence in economics, rather than psychology, it was sometimes implicitly assumed that rewards are purely self-interested in the narrowest sense: the food one eats, the money one earns, the social approbation one craves, and so on. Of course nobody ever suggested that people fail to act in ways that benefit others—people obviously do each other favors, give each other gifts, pay each other complements, and so on. Rather, it was sometimes implicitly assumed that people perform these apparently generous actions not because they provide us with intrinsic reward, but instead because they are instrumentally valuable. That is, we do nice things for other people just because eventually it helps us eat, earn, and bask in glory.

Against this background, several decades of behavior studies have been devoted to demonstrating that people experience a variety of social events and outcomes as intrinsically rewarding: Fairness (Fehr & Schmidt 1999), giving (Andreoni 1990; Crumpler & Grossman 2008), reciprocity (Rabin 1993), revenge (Carlsmith et al. 2002; Morris et al. 2017) and norm enforcement (Fehr & Gächter 2002; Fehr & Fischbacher 2004), conformity to norms (Capraro & Rand 2018), and much more. The standard experimental approach is to devise an experiment (usually in a laboratory and for monetary stakes) in which a person has the opportunity to pay a cost to attain the good in question (fairness, revenge, etc.), but with the knowledge that their behavior will be anonymous and the interaction "one shot". This removes any possibility of recouping their losses, and indicates that the value of their behavior must be derived from an intrinsic sense of reward.

This logic is compelling if we assume that peoples' behaviors in the experiments are derived from representations of value. But, as we have already noted, there are other kinds of psychological mechanisms that might explain their behavior without any representation of value: They might be behaving reflexively, or recapitulating social scripts or norms, or repeating the sorts of behaviors they've performed in the past, etc., without computing the expected value of candidate actions and choosing between them on that basis. A crucial complementary piece of evidence comes from cognitive neuroscience. In case after case, the neural systems and neural signatures associated with social decision-making turn out to be the very same ones that have long been associated with value-based decision-making

more generally (reviewed in Ruff & Fehr 2014; Rilling & Sanfey 2011). Decades of work have shown how neural populations within the striatum, ventromedial prefrontal cortex, orbitofrontal cortex, and elsewhere implement a well-characterized set of computations in order to estimate the value of objects, states, and actions in order to guide choice (reviewed in Rangel et al. 2008). And, these very same circuits are involved in making decisions about fairness (Sanfey et al. 2003; Hsu et al. 2008), giving (Zaki & Mitchell 2011), reciprocity (Phan et al. 2010; Van den Bos et al. 2009), punishment (Treadway et al. 2014; Sanfey et al. 2003), spite (Cikara et al. 2011), conformity (Zaki et al. 2011; Izuma & Adolphs 2013), reputation (Izuma et al. 2008) and much more. Taken together, then, the neural evidence suggests that value-based mechanisms play a key role in social decision-making, while behavioral studies indicate that this is often because we assign intrinsic rewards to social actions and events.

Of course, there are also many instances in which our social behavior is guided by instrumental reasoning—where we strategically derive the value of actions, states, or even people, by considering their implications for our own "self-interested" rewards. For instance, we assign value to generous actions strategically based upon how others will perceive them (Izuma et al. 2010; van Baar et al. 2019). We also assign value to individuals based on their propensity to act generously towards ourselves and others (Behrens et al. 2008; Hackel et al. 2015).

In retrospect, it should not be surprising that people experience social rewards and assign social values, or that these rely on common computational schemes and, in many cases, common neural substrates with non social rewards and values. Nevertheless, this fact has driven tremendous progress in modeling the computational basis of social decision-making. After all, the study of value-guided decision-making has been one of the great success stories in human behavioral research—a rare instance where computational, behavioral , and neural studies have converged on a common core set of models.

4.2. Estimating value: Learning and planning

In many cases it can be extremely challenging to estimate the value of an action. Consider, for instance, the problem of choosing a chess move. Although reward is well-defined in this setting (checkmate), determining which actions are most likely to *bring about* that reward is difficult to compute. Part of the problem is to model the other player, which we consider in the following section on multi-agent decision-making and game theory. Here, we focus on another part of the problem, essential to what is sometimes called the "reinforcement learning" (RL) problem (Sutton & Barto 2018), which is to choose adaptive *sequences* of behavior. In computer science, neuroscience, and psychology, considerable progress has been made in understanding the different sorts of strategies one can use to estimate value in sequential settings. These are playing an increasingly prominent role in social psychology.

Reinforcement learning: A family of methods that select sequences of actions by estimating their values.

4.2.1. Model-based versus model-free methods. Upon first thought, it might seem straightforward to estimate the value of an action like a chess move: One merely needs to consider the resulting position of the board, the probabilities of various moves by the opponent, the probabilities of your replies, and so on, until each imagined chain ends in checkmate. This is often called planning, and it is an example of a "model-based" method of value estimation, because it depends on one's probabilistic causal model of how various actions are related to subsequent states. In this case, the model both involves knowledge of the (relatively

Model-based: Methods of estimating value that draw on a causal model linking actions to outcomes, such as planning.

Model-free: Methods of estimating value without drawing on a causal model, such as generalizing from past experience.

simple) rules of chess, and also knowledge of the (relatively complicated) decision-making processes of one's opponent (i.e., precisely the kind of causal model discussed in Section 2: A model of the rational actor). But, the same basic computation can be employed for any decision problem for which you have a causal model: choosing a path based on a model of the terrain; choosing a repair based on a model of one's bicycle; choosing a house based on a model of one's future needs; etc. For a computationally unbounded agent, model based methods can provide an optimal approach to value estimation. In practice, however, they are often prohibitively computationally demanding. Thus, a growing body of research suggests that humans engage in planning over simplified task models (Ho et al. 2022a), limited time horizons (Keramati et al. 2016), restricted sets of goals or options (Morris et al. 2021; Cushman & Morris 2015), or in other ways designed to make the planning problem tractable.

An alternative way to estimate the value of an action is according to its history of past reward. For instance, the value of a particular chess move in the present game could be approximated by tallying the proportion of times it was a winning move in prior games. This does not require building or querying a causal model of chess games, but instead keeping statistics on past performance. It is an example of a "model-free" method of value estimation. An especially efficient and powerful model-free method, temporal difference learning (Watkins & Dayan 1992; Sutton 1988), exploits an idea closely related to the "secondary reinforcer" (i.e., an event that is treated as rewarding not because it carries intrinsic reward, but because it has historically *predicted* subsequent rewards). Each time one takes an action, one updates the learned value of that action by adjusting it in the direction of two quantities, summed: the *immediate* rewards one experiences in the next state, and one's estimate of the value of the *next* action one will take in that new state. This method efficiently "passes back" information about expected future rewards through a simple update rule over the course of repeated trial-and-error experience.

Early model-free algorithms required that agents gradually improve their estimates of the value of various actions only given experience with precisely those actions in precisely those states. Yet there is some sense in which people actually *never* encounter precisely the same state twice—and, in any event, a more efficient approach is to generalize across experience based on the overall similarity between the current action and state and previous ones that have been encountered. This approach is common to many "deep RL" methods employed in contemporary AI architectures, where a deep neural net approximates a value function defined over states and actions through statistical generalization (e.g. Silver et al. 2017). In architectures of this kind, the process of generalization is generally model-free.

Compared with model-based methods, model-free methods tend to require greater quantities of training and do not make optimal use of evidence, so they tend to perform less accurately. Their advantage, however, is computational efficiency at decision time. Whereas model-based planning requires an agent to prospect over a potentially large and lengthy set of subsequent states and actions, model-free methods typically require simple lookup of a precompiled (or "cached") representation of value, or a simple query of a precompiled value function.

There is a long and productive tradition of "dual process" models in social psychology, as well as in judgment and decision-making, cognitive psychology, and behavioral economics (Kahneman 2011; Sloman 1996). Broadly speaking, these models posit that we can accomplish many tasks either by devoting conscious, effortful and controlled cognitive processes to deriving a more accurate solution, or instead by relying on unconscious, effortless and

automatic processes that deliver less accurate approximations. The distinction between model-free and model-based RL offers one of the most compelling and useful computational formalizations of this basic idea, albeit just for the one psychological domain of value-guided decision-making.

For instance, the dual process model of moral judgment posits that we have an automatic process of moral judgment that prohibits direct physical harm to others, while we have a controlled process that endorses whichever action maximizes the long-run welfare of those we care about (Greene 2008). These processes can conflict in cases like the trolley problem, where one brings about the greatest long-run welfare by directly harming a person. This model can be naturally interpreted within the RL framework as a competition between a model-free process that assigns low value to harming others (because it is usually a bad choice) and a model-based process that is able to derive the situation-specific *high* value of harming others in unusual cases like the trolley problem (Cushman 2013; Crockett 2013). This formalization of the dual process of morality clarified its relationship to other work on value-guided decision-making, while also generating a host of new predictions that have begun to receive empirical support (Patil et al. 2021; Lockwood et al. 2020; Maier et al. 2023).

4.2.2. Individual versus social learning methods. Social psychology has much to gain from models of value-guided decision-making, but the reverse is true as well. One of the strangest features of the contemporary literature on value-guided decision-making is its overwhelming focus on individual learning. That is, it is often assumed that people learn about the causal structure of the world, and the value of actions and states within it, through personal experience. Yet, surely the cornerstone of humans' distinctive intelligence is our capacity to learn from and teach each other (Tomasello 2009). Over the past decade, increasing attention has been paid to developing formal accounts of the intersection between social learning and value-based decision-making (Olsson & Phelps 2007; Najar et al. 2020; Vélez & Gweon 2021; Charpentier et al. 2020; Wu et al. 2022; Morris & Cushman 2018).

A foundational question in the study of human social learning, long recognized by developmental psychologists and anthropologists, is the extent to which we learn by blindly imitating others' successful actions, or instead by developing a causal understanding of what makes those actions successful and then deriving suitable plans for ourselves given that knowledge (Boyd et al. 2011; Pinker 2010; Tomasello 2009). These two forms of social learning can be considered analogous to model-free and model-based methods of individual learning: in the former one copies one's own prior successful actions, while in the latter one learns a causal model and then flexibility plans in future circumstances given this knowledge (Charpentier et al. 2020). Thus, computational models of value-guided decision-making offer a powerful and appealing way of organizing theories of social learning (Najar et al. 2020; Wu et al. 2022; Morris & Cushman 2018).

Several formal models seek to define social learning strategies that combine elements of Bayesian mental state inference and value-guided decision-making (Charpentier et al. 2020; Najar et al. 2020; Ho et al. 2021, 2019; Thompson et al. 2022). In other words, when attempting to learn from an expert, it can be helpful to infer the mental states (rewards, values, and beliefs) that guide their behavior—a classic instance of Bayesian mental state inference—and then to use those inferred mental states to appropriately adjust and improve your own rewards, values, and beliefs. And, when teaching, it helps to balance the cognitive and physical cost of demonstration against the epistemic benefit to the learner (Vélez et al.

2023) This line of work highlights the appealing compositionality of computational cognitive models.

4.3. Value-based Choice

So far we have considered how people learn a representation of value. A separate question is how, given such a representation, a person makes their choice.

Consideration set: In open-ended decision problems, the restricted set of options, or actions, subject to deliberation.

4.3.1. Consideration. Many laboratory studies of choice present participants with exactly two alternatives. But the real world often presents "open-ended" decision problems with far too many alternatives to explicitly consider them all: Too many meals we could possibly eat, too many moves in a game of Go, too many things we could do on a Sunday afternoon. When people face open-ended decisions, a limited set of good candidate options tends to come readily to mind, often called the "consideration set". Research in consumer behavior suggests that composition of the consideration set often explains as much or more about a person's ultimate choice than the deliberative processes by which they select a final item from within the consideration set. Recently, considerable progress has been made in formalizing the processes by which the consideration set is constructed (Morris et al. 2021; Aka & Bhatia 2021; Zhang et al. 2021; Callaway et al. 2022), focusing both on the role of semantic associations (e.g. the associability of different foods with "dinner") and of value representations (e.g., the value associated with different foods as dinner items).

Although still circumstantial, there is some evidence that social norms—and especially moral values—guide consideration (Phillips & Cushman 2017; Bernhard et al. 2022; Phillips et al. 2019; Swidler 1986; Kalkstein et al. 2022). In other words, when choosing our own actions, the candidates that come naturally to mind are those that others typically perform, and that do not violate moral norms. For instance, if we are on the way to the airport, our car breaks down, and we're out of cash, we tend to consider possibilities like calling a friend for a ride (common, and morally acceptable) rather than possibilities like calling a cab and simply not paying for it when we arrive at the airport (uncommon, and morally wrong). An important direction for future research is to embed this role of social and moral norms into formal accounts of consideration (Morris & Cushman 2018).

4.3.2. Evidence accumulation. Once a set of candidate actions is under consideration, how do we deliberate about their relative values? Typically we cannot query a single deterministic representation of value, but instead draw "samples" from a generative process. The more samples we draw, the more accurate our estimate. For instance, if estimating the value of chess move, we might be able to consider more potential future states of the board; if estimating the value of a home, we might be able to consider more of its features; if we are estimating the value of a friendship, the more past episodes we might be able to recall from memory, and so on. Because we cannot deliberate over the options in our consideration set forever, however, people often employ a confidence threshold for choice.

A broad class of "accumulator" models provide a formal description of this process (Ratcliff & Smith 2004). The most popular of these, the drift diffusion model (DDM), proposes that we sample the relative value of options until we reach a threshold surplus of samples favoring one of the options. In other words, the DDM tells you not only what to choose, but also when you've devoted enough time and cognitive effort to make a choice with some degree of certainty. Originally developed to model perceptual tasks (e.g., are most

of these moving dots going right or left?), it provides an excellent fit to both judgment and reaction time in many contexts (Krajbich et al. 2010; Fudenberg et al. 2020), and also presents intriguing parallels to neural data on the choice process (Gold & Shadlen 2007; Harris & Hutcherson 2022).

Social psychologists have made several productive uses of the DDM. First, it has illuminated specific mechanisms by which social factors affect decision processes. For instance, a DDM applied to the effect of conformity on moral decision-making shows that it does not to directly bias people match a particular outcome, but instead to bias their attention to the choice attributes that align with their peers' decision-making (Yu et al. 2021). (In contrast, reward-motivated biases in choice behavior show an influence of both kinds; see Leong et al. 2019). Another shows that peer influence increases impulsivity during decision-making (Son et al. 2019).

Second, by permitting quantitative modeling of reaction time effects, the DDM has allowed researchers to question some claims of "dual-process" models based on the observation that people make certain decisions faster than others. Thus, for instance, although people may tend to choose unhealthy snacks when deciding quickly and healthy snacks when taking their time, a DDM indicates that this may simply reflect that they are able to access information about taste sooner than health, rather than a competition between distinct automatic and controlled processes (Sullivan et al. 2015). Similarly, evidence that people make prosocial decisions more quickly than pro-self decisions has been used to argue that prosocial behavior is generated by automatic psychological processes while pro-self behavior is generated via controlled deliberation (Rand et al. 2014, 2012). Using a DDM, however, it has been argued that a similar pattern of data could instead be generated by a single system (Hutcherson et al. 2015).

Third, researchers have used the DDM to explore how ordinary people draw mental state inferences based on others' decision time. For instance, if you ask somebody to marry you and they take a heartbeat versus a minute to answer, you might draw very different conclusions about their level of enthusiasm and certainty (Critcher et al. 2013; Pizarro et al. 2003)—potentially the same sorts of conclusions that a computational cognitive scientist would draw by using a DDM. Indeed, ordinary people's inferences about others' preferences based on reaction time align with the predictions of a "rational" model that uses Bayes' rule to perform inference over a DDM (Gates et al. 2021).

This final study showcases the compositionality of formal methods: the DDM (devised for perceptual decisions) can be composed with theories of value guided decision-making, which can themselves be composed with models of Bayesian inference, to model how people draw inferences about others values based on reaction time. But, it illustrates how formal models are sometimes best construed as hypotheses about how psychological processes are *implemented* (at Marr's implementation or algorithmic levels), and other times best construed as hypotheses about how psychological proceses are *designed* (at Marr's computational level). When people make value-guided decisions, it seems that they are often implementing a mechanism very much like the DDM. On the other hand, when people draw inferences about others' values based on reaction time, it is quite unlikely that they are performing exact Bayesian inference over a generative *model* of the DDM. Nevertheless, the same formal tools can help us to understand the design principles underlying their inference.

4.3.3. Rationalization. Perhaps peculiarly, the processes of value-guided decision-making does not end with choice. A classic area of research in social psychology is rationalization: Our tendency to construct rational explanations of our behavior *post-hoc*. One of the most robust empirical findings in this literature is choice-induced preference change: i.e., the phenomenon that choosing things tends to make us like them more, and failing to choose them tends to make us like them less (e.g., Brehm 1956; Sharot et al. 2009; Koster et al. 2015). This, of course, turns the standard model of value-guided choice—in which one chooses something *because* it is valuable—on its head.

A variety of computational accounts of choice-induced preference change have been offered at the mechanistic level (e.g. Vinckier et al. 2019; Cockburn et al. 2014). A separate question arises, however, as to why such a mechanism would be advantageous. What adaptive purpose is served by having people tend to prefer actions *because* they chose them, and not just the other way around? One possibility is that this reflects a process of trying to achieve coherence between various parts of one's psychology, including beliefs, values, and actions (Hornsby & Love 2020; Holyoak & Simon 1999; Shultz & Lepper 1999).

Another, related, possibility is that choice-induced preference change serves the function of *inferring* the value of objects given choice. This may be because people do not have introspective access to their choice processes and yet wish to be able to explain their behavior convincingly (and perhaps flatteringly) to themselves and others (Bem 1972; Von Hippel & Trivers 2011). It may also be because rationalization of this kind can improve one's reasoning and subsequent decision-making (Cushman 2020). People sometimes perform behaviors that are not guided by an explicit representation of reward (e.g., instinctual behaviors, conformity to social norms, certain habitual actions, etc.), yet which are nevertheless adaptive. Thus, by inverting a generative model of rational action (along the lines of the Bayesian models of theory of mind considered in Section 2), an agent can discover what they ought to find rewarding—for instance, concluding that certain foods are nourishing by rationalizing innate taste preferences.

5. MULTI-AGENT CHOICE

So far we have addressed value-based decision-making in which a person interacts principally with their environment, but of course special considerations arise when multiple agents make choices that influence each other's behavior and outcomes.

There is a long history of using formal models to explain how people influence each other's behavior. For instance, following Asch's (1955) seminal characterization of the interplay of individual judgment and conformity, the Bass diffusion model successfully formalized how consumers' adoption of new products are influenced by conformity (Bass 1969). In quite a different domain, Coleman & James (1961) provided a computational analysis of the expected size of crowds based on their inflows and outflows, which Milgram et al. (1969) assessed against experimental data. A particularly influential model due to Schelling (1971) shows how even small preferences to live a low threshold of people similar to us, once aggregated and played out over time, can result in a profoundly segregated population.

Each of these models draws powerful, abstract insights from simple grounding assumptions. Yet, each is grounded in a rather *ad hoc* setup, which restricts their compositional potential. What follows is a selective review of game-theoretic approaches to multi-agent behavior as class of models that compose naturally with the other formal frameworks we have reviewed (i.e., Bayesian inference and value-based decision-making).

5.1. Game theory

Equilibrium: A set of strategies, implemented in a multi-agent game, such that no single agent can unilaterally improve their expected payoff by changing their strategy.

Game theory models multi-agent interactions by drawing on two foundational ideas that we have already encountered. First, it assumes that people make rational decisions aimed at maximizing value—a claim about decision-making. Second, it assumes that they model *others* as rational value-maximizers as well—a claim about inference and prediction.

One of the most essential ideas arising from this formal approach is the "equilibrium": a set of strategies (one for each player in the game) which, when simultaneously implemented, prevent any player from improving their expected payoff unilaterally[2]. For instance, if we regard driving as a game, one equilibrium is for all drivers to always drive on their right-hand side of the road, and another equilibrium is for all drivers to always drive on their left-hand side. In either case, nobody can do better by changing their behavior alone. In contrast, the strategy "drive on the right with 70% probability" is not an equilibrium: if everybody behaved this way, anybody could improve their payoff by driving on the right side *more* often (thus reducing their probability of an accident). Rational actors would do so, bringing the population in line with one of its true equilibria: Everyone driving on the right. This is what makes the equilibrium a powerful concept: it predicts which behaviors we should tend to observe in multi-agent interactions, and explains why.

In many different areas of interest to social psychologists, the kinds of behaviors that we observe in the real world can be elegantly explained as equilibrium solutions to simple games. This is in part because many different adaptive processes are all predicted to converge to equilibria: Model-based reasoning (e.g. Nash 1951), model-free reinforcement (e.g. Littman 2001), natural selection operating on genetically inherited behavior (e.g. Smith & Price 1973), and cultural selection operating on socially learned behavior (e.g. Boyd & Richerson 1988). Thus, game theory can organize our understanding of human interaction even without specifying the precise mechanisms involved. Of course, as such, this relegates many applications of game theory to a position at Marr's (1982) computational, or Tinbergen's (1963) adaptive, level of analysis.

5.1.1. Signaling. Many game theoretic models ask when communication is reliable: In other words, when people will speak honestly, and trust what they hear. It has been long been recognized that *costly* signals might be used to establish trust. Veblen (1889) posited, for instance, that the rich could "burn money" on objects (or time, on activities) with little utility in order to send a credible signal of their resources and power. What makes the signal credible, on this analysis, is that a poor person would not rationally choose to pay it. This intuition was later formalized in a series of game theoretic models outlining the conditions in which costly signals can establish successful communication, the conditions under which costs are not necessary, and the conditions under which dishonesty proliferates despite costs (Grafen 1990; Smith 1994, 1991).

Costly signals can be used to establish more than just one's wealth and power, of course. For instance, they can signal one's commitment as a romantic partner (Zahavi 1975), such as when a person pays for an expensive meal on a date. Or, they can be used to signal the importance of a ritual or cultural practice, such as when extravagant expenses are thrown to observe holidays or honor gods (Henrich 2009). And, in one of the best-

[2]There are numerous different equilibrium concepts that are important to game theory; this statement is meant to capture their shared foundation.

explored examples, they can be used to signal one's value as a cooperative partner (Gintis et al. 2001). That is, by engaging in particularly costly acts of cooperation or altruism, one can send an honest and reliable signal of the importance of one's moral reputation and, by extension, social relationships. Remarkably, formal analysis shows that one can send precisely the same kind of signal of trustworthiness as a social partner by engaging in costly third-party *punishment* of norm violators, and empirical data suggest this is in fact an important motivation for third-party punishment in some cases (Jordan et al. 2016a). Taken together, these models have been used to explain phenomena like virtue signaling and competitive altruism (Roberts 1998; Barclay & Willer 2007; Kraft-Todd et al. 2020).

5.1.2. Emotion, intuition, and commitment. In some games, paradoxically, we can achieve our best outcomes by committing ourselves to irrational behavior. The classic example is the doctrine of mutual assured destruction in nuclear war, a strategic setting formalized as a game by Schelling (1960). Although it may be irrational to launch a counter-strike once one's destruction is already assured, *pre-committing* to the counter-strike can act as an important deterrent against would-be aggressors. In other words, by committing to irrational retaliation, one can rationally prevent an initial attack. Similarly, when people possess valuable resources that can easily be stolen—perhaps cattle, drugs, or clients—it can be strategically beneficial for them to advertise an "irrational" commitment to massive retaliation against anybody who threatens those resources. Certain moral emotions such as a sense of "honor", or vengeance, are suggested to accomplish this function (Frank 1988): When we exhibit these emotions, we send a credible signal to people around us that we will retaliate against any offense, and even at great personal cost. This deters would-be offenders. This logic is essential to the standard anthropological and social-psychological analysis of honor cultures, such as the herders and ranchers of the American South and West (Nisbett 2018), and elsewhere (Cao et al. 2021).

Frank (1988) argued that many emotions, not just anger and revenge, could be analyzed as commitment devices. He proposed that romantic love, for instance, plays the role of "irrationally" committing us to fidelity which, in turn, enables the profitable institution of highly trusting long-term romantic relationships. Similar arguments, grounded in game theory, have been applied to a variety of other emotions and other inflexible, evolved motivational systems such as intuitions and social rewards (Hoffman et al. 2015; Jordan et al. 2016b; Bear & Rand 2016; Morris et al. 2017; Bernhard & Cushman 2022).

5.1.3. Fairness: Bargaining and coordination. One of the earliest and most successful applications of game theory was to cooperative bargaining (Nash 1950): A situation in which two people can achieve mutual benefit by working together to create value, as long as they can strike a deal over how to divide the profits. This kind of human interaction is ubiquitous: Our relationships with employers, romantic partners, friends, businesses, neighborhood associations, and so forth, all involve collaborations in which people work together for mutual benefit, and must find a way to share the profits. Formal analysis shows that rational agents will tend to divide profits by assigning the greater share to an individual with superior "outside options"—i.e., the person who does better if no deal is struck. In large social markets, however—i.e., situations in which there are many potential social partners with whom a person might interact, and thus a wealth of outside options for all—rational agents will tend to offer an even (50/50) division of resources in order to maintain their competitive value as a social partner (André & Baumard 2011). This has been proposed to account for

a broad array of human moral norms, especially concerning distributive justice (Baumard et al. 2013; Binmore 2005; Le Pargneux et al. 2023).

In real-world contexts, however, coordinating our behavior in order to achieve mutual benefit can be challenging. Consider the mundane complexity of a married couple sorting out who will handle various household tasks: shopping, cooking, cleaning, home maintenance, pet walking, childcare, and so on. In these contexts, coordination can become more efficient and reliable by resorting to conventionalized social roles, such as gendered expectations, but at the price of inequality. Formal analysis shows that not just our intuitions about fairness, but also the persistence of unfairness, can be elegantly captured in the language of game theory (O'Connor 2019).

5.1.4. Reciprocity. No facet of human sociality has been more extensively modeled as games than reciprocity in social dilemmas: Our tendency to cooperate with those who are also cooperative, and to withhold cooperation (or otherwise punish) those who are not. In games like the two-player Prisoner's Dilemma (PD) individuals maximize their immediate payoff by withholding cooperation, but achieve a higher payoff through mutual cooperation than through mutual defection—hence the "dilemma". Classic formal models (Trivers 1971) and agent-based simulations (Axelrod & Hamilton 1981) showed that in certain *repeated* PDs, however, cooperation can emerge from an equilibrium in which each player cooperates only with those who have previously been cooperative. Subsequent work revealed additional equilibria that explain, for instance, why people cooperate with those who have a cooperative *reputation*, or why especially strong cooperative norms might be maintained within (but not between) small, close-knit groups that compete (Nowak 2006).

Many human emotions scaffold reciprocity in social dilemmas (Sznycer 2019; Chang & Smith 2015). When people act less cooperatively than expected we feel anger and resentment (Sell 2011; Sell et al. 2017). Meanwhile, when we detect the anger and resentment of others we are often motivated to make amends (Sznycer et al. 2016; Chang et al. 2011). Often, in order to adequately model the structure of human moral emotions in game theoretic terms, one must explicitly model each player's *beliefs* about the other player, including their beliefs about the other players' knowledge or intentions (Rabin 1993; Chang et al. 2011). The method of integrating models of mental state inference (such as those described in Section 2) with game theory has been a fertile area of work for several decades (Geanakoplos et al. 1989; Battigalli & Dufwenberg 2009).

5.1.5. Norms and equilibrium selection. Norms are enforced behavioral standards. Several different formal models of norms share a similar logic. The norm dictates that one must behave a certain way (contributing to a public food, for instance) or else be in bad standing, and it dictates that one must punish (or withhold cooperative benefits from) those in bad standing, or else be in bad standing (Bicchieri 1990; Ostrom 2000; Boyd & Richerson 1992). The self-reinforcing nature of this arrangement is quite apparent: Not only the norm is enforced, but also enforcement is enforced. Indeed, under certain assumptions, this logic can render *any* behavioral standard an equilibrium, even if the behavior itself is contrary to peoples' interests (Boyd & Richerson 1992). Thus, for instance, in many societies there is a norm that one must contribute to public goods (which creates welfare and produces mutual benefit), but in other societies there is a norm that one ought *not* to contribute to public goods—those cooperators who contribute thus get *punished* (Herrmann et al. 2008).

At first blush, the power of self-enforcing social norms to explain the maintenance of *any*

Equilibrium selection: A process by which certain equilibria come to predominate over others in a game.

arbitrary behavior would seem to make a mockery of the predictive power of the equilibrium concept—after all, *anything* could be predicted. In order to help us predict which specific equilibria will tend to be enforced by social norms—and, therefore, to explain the social norms we tend to see around us—we need the additional conceptual tool of "equilibrium selection" (Harsanyi et al. 1988). As noted above, different equilibria will yield different payoffs: A social norm that enforces the equilibrium "do contribute to public goods" will tend to leave people better off, for instance, than one enforces the equilibrium "don't". (Unsurprisingly, countries characterized by the former norm tend to be wealthier than those characterized by the latter). Because people prefer social arrangements that leave them better off, they will sometimes choose social norms that they anticipate will lead to better equilibria (Levine et al. 2020; Redish 2022), and both cultural and biological evolutionary processes also tend to favor those equilibria that produce better (i.e., fitter) outcomes (Henrich 2004; Boyd & Richerson 1992).

6. CONCLUSION

Social psychology is animated by the impulse to bring scientific order to the complexity of human social life through basic principles of mental organization. Surveying the state of their fledgling field, it founders saw considerable energy devoted to documenting the complexity, but relatively little to ascertaining the underlying principles. They assumed, however, that formal methods would eventually reveal the hidden logic of human sociality.

This hope is no less ambitious today than it was 80 years ago, yet we have made considerable progress. Across the social cognitive sciences, a family of related models posits that we represent the world, and each other, largely in terms of probabilistic causal models. They explain how we can learn much from limited data, by the logic of Bayesian inference. They explain how we organize our knowledge into simple and practical attributions of responsibility. They explain how we can make adaptive decisions, via the logic of reward and value. And, they explain how we negotiate the interdependence of our choices, via the logic of game theory. Each of these theories can be fruitfully composed with the others. The same probabilistic causal models support attribution, inference, and choice. The same logic of reward maximization supports people's own choices and their inferences about how others will choose. And, rational choices of this kind—together with a mental model of others' rational choices—will tend to lead human agents towards the equilibria predicted by game theory.

Any serious social scientist must occasionally experience crises of confidence about our collective intellectual enterprise. Human sociality is confounding complex, and it can seem that every claim is subject to counterexamples and contextual limitations. Ideas are chronically misunderstood or misconstrued. We squander precious years adjudicating the boundaries of our theories, their predictions, and the very meaning of the concepts they invoke.

It is against this background that formal models offer their greatest appeal. Because of their compositionality, the conventions upon which they depend, and the level of abstraction at which they are defined, it becomes clearer what each hypothesis states, and what each concept means; how it differs from others; how data would bear on it; how it adds to our collective state of knowledge; how it can be applied and extended beyond its original purview. Each individual contribution remains small, but numbers have a way of adding up. Eventually it becomes apparent that we are, after all, collaborating on a project of ever-growing explanatory reach.

ACKNOWLEDGMENTS

Thanks to the Moral Psychology Research Laboratory for helpful discussions and feedback. This work was supported by grant N00014-22-1-2205 from the Office of Naval Research.

LITERATURE CITED

Aka A, Bhatia S. 2021. What i like is what i remember: Memory modulation and preferential choice. *Journal of Experimental Psychology: General* 150:2175

Alicke MD. 2000. Culpable control and the psychology of blame. *Psychological bulletin* 126:556

André JB, Baumard N. 2011. The evolution of fairness in a biological market. *Evolution* 65:1447–1456

Andreoni J. 1990. Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal* 100:464–477

Asch SE. 1955. Opinions and social pressure. *Scientific American* 193:31–35

Axelrod R, Hamilton WD. 1981. The evolution of cooperation. *science* 211:1390–1396

Baker CL, Jara-Ettinger J, Saxe R, Tenenbaum JB. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1:1–10

Baker CL, Saxe R, Tenenbaum JB. 2009. Action understanding as inverse planning. *Cognition* 113:329–349

Barclay P, Willer R. 2007. Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences* 274:749–753

Bass FM. 1969. A new product growth for model consumer durables. *Management science* 15:215–227

Battigalli P, Dufwenberg M. 2009. Dynamic psychological games. *Journal of Economic Theory* 144:1–35

Baumard N, André JB, Sperber D. 2013. A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences* 36:59–78

Bear A, Rand DG. 2016. Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences* 113:936–941

Behrens TE, Hunt LT, Woolrich MW, Rushworth MF. 2008. Associative learning of social value. *Nature* 456:245–249

Bem DJ. 1972. Self-perception theory. In *Advances in experimental social psychology*, vol. 6. Elsevier, 1–62

Bernhard RM, Cushman F. 2022. Extortion, intuition, and the dark side of reciprocity. *Cognition* 228:105215

Bernhard RM, LeBaron H, Phillips J. 2022. It's not what you did, it's what you could have done. *Cognition* 228:105222

Bhui R, Lai L, Gershman SJ. 2021. Resource-rational decision making. *Current Opinion in Behavioral Sciences* 41:15–21

Bicchieri C. 1990. Norms of cooperation. *Ethics* 100:838–861

Binmore K. 2005. Natural justice. Oxford university press

Boyd R, Richerson PJ. 1988. Culture and the evolutionary process. University of Chicago press

Boyd R, Richerson PJ. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and sociobiology* 13:171–195

Boyd R, Richerson PJ, Henrich J. 2011. The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences* 108:10918–10925

Brehm JW. 1956. Postdecision changes in the desirability of alternatives. *The Journal of Abnormal and Social Psychology* 52:384

Callaway F, van Opheusden B, Gul S, Das P, Krueger PM, et al. 2022. Rational use of cognitive resources in human planning. *Nature Human Behaviour* 6:1112–1125

Cao J, Kleiman-Weiner M, Banaji MR. 2017. Statistically inaccurate and morally unfair judgements via base rate intrusion. *Nature Human Behaviour* 1:738–742

Cao J, Kleiman-Weiner M, Banaji MR. 2019. People make the same bayesian judgment they criticize in others. *Psychological Science* 30:20–31

Cao Y, Enke B, Falk A, Giuliano P, Nunn N. 2021. Herding, warfare, and a culture of honor: Global evidence. Tech. rep., National Bureau of Economic Research

Capraro V, Rand DG. 2018. Do the right thing: Experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Judgment and Decision Making* 13:99–111

Carlsmith KM, Darley JM, Robinson PH. 2002. Why do we punish? deterrence and just deserts as motives for punishment. *Journal of personality and social psychology* 83:284

Chang LJ, Smith A. 2015. Social emotions and psychological games. *Current Opinion in Behavioral Sciences* 5:133–140

Chang LJ, Smith A, Dufwenberg M, Sanfey AG. 2011. Triangulating the neural, psychological, and economic bases of guilt aversion. *Neuron* 70:560–572

Charpentier CJ, Iigaya K, O'Doherty JP. 2020. A neuro-computational account of arbitration between choice imitation and goal emulation during human observational learning. *Neuron* 106:687–699

Cikara M, Botvinick MM, Fiske ST. 2011. Us versus them: Social identity shapes neural responses to intergroup competition and harm. *Psychological science* 22:306–313

Cockburn J, Collins AG, Frank MJ. 2014. A reinforcement learning mechanism responsible for the valuation of free choice. *Neuron* 83:551–557

Coleman JS, James J. 1961. The equilibrium size distribution of freely-forming groups. *Sociometry* 24:36–45

Critcher CR, Inbar Y, Pizarro DA. 2013. How quick decisions illuminate moral character. *Social Psychological and Personality Science* 4:308–315

Crockett MJ. 2013. Models of morality. *Trends in cognitive sciences* 17:363–366

Crockett MJ, Everett JA, Gill M, Siegel JZ. 2021. The relational logic of moral inference. In *Advances in Experimental Social Psychology*, vol. 64. Elsevier, 1–64

Crumpler H, Grossman PJ. 2008. An experimental test of warm glow giving. *Journal of public Economics* 92:1011–1021

Cushman F. 2008. Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108:353–380

Cushman F. 2013. Action, outcome, and value: A dual-system framework for morality. *Personality and social psychology review* 17:273–292

Cushman F. 2020. Rationalization is rational. *Behavioral and Brain Sciences* 43:e28

Cushman F, Macindoe O. 2009. The coevolution of punishment and prosociality among learning agents, In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 31

Cushman F, Morris A. 2015. Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences* 112:13817–13822

Diaconescu AO, Mathys C, Weber LA, Daunizeau J, Kasper L, et al. 2014. Inferring on the intentions of others by hierarchical bayesian learning. *PLoS computational biology* 10:e1003810

Diaconescu AO, Mathys C, Weber LA, Kasper L, Mauer J, Stephan KE. 2017. Hierarchical prediction errors in midbrain and septum during social learning. *Social cognitive and affective neuroscience* 12:618–634

Fehr E, Fischbacher U. 2004. Third-party punishment and social norms. *Evolution and human behavior* 25:63–87

Fehr E, Gächter S. 2002. Altruistic punishment in humans. *Nature* 415:137–140

Fehr E, Schmidt KM. 1999. A theory of fairness, competition, and cooperation. *The quarterly journal of economics* 114:817–868

FeldmanHall O, Shenhav A. 2019. Resolving uncertainty in a social world. *Nature human behaviour*

3:426–435
Frank RH. 1988. Passions within reason: The strategic role of the emotions. WW Norton & Co
Fudenberg D, Newey W, Strack P, Strzalecki T. 2020. Testing the drift-diffusion model. *Proceedings of the National Academy of Sciences* 117:33141–33148
Gates V, Callaway F, Ho MK, Griffiths TL. 2021. A rational model of people's inferences about others' preferences based on response times. *Cognition* 217:104885
Geanakoplos J, Pearce D, Stacchetti E. 1989. Psychological games and sequential rationality. *Games and economic Behavior* 1:60–79
Gershman SJ. 2021. The rational analysis of memory. *Oxford handbook of human memory.*
Gershman SJ, Pouncy HT, Gweon H. 2017. Learning the structure of social influence. *Cognitive Science* 41:545–575
Gerstenberg T, Goodman ND, Lagnado DA, Tenenbaum JB. 2021. A counterfactual simulation model of causal judgments for physical events. *Psychological Review* 128:936
Gintis H, Smith EA, Bowles S. 2001. Costly signaling and cooperation. *Journal of theoretical biology* 213:103–119
Gold JI, Shadlen MN. 2007. The neural basis of decision making. *Annu. Rev. Neurosci.* 30:535–574
Goodman ND, Frank MC. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences* 20:818–829
Grafen A. 1990. Biological signals as handicaps. *Journal of theoretical biology* 144:517–546
Greene JD. 2008. The secret joke of kant's soul. *Moral psychology* 3:35–79
Griffiths TL, Kemp C, Tenenbaum JB. 2008. Bayesian models of cognition
Gweon H. 2021. Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences* 25:896–910
Hackel LM, Doll BB, Amodio DM. 2015. Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice. *Nature Neuroscience* 18:1233–1235
Halpern J, Kleiman-Weiner M. 2018. Towards formal definitions of blameworthiness, intention, and moral responsibility, In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32
Halpern JY, Hitchcock C. 2015. Graded causation and defaults. *The British Journal for the Philosophy of Science*
Harris A, Hutcherson CA. 2022. Temporal dynamics of decision making: A synthesis of computational and neurophysiological approaches. *Wiley Interdisciplinary Reviews: Cognitive Science* 13:e1586
Harsanyi JC, Selten R, et al. 1988. A general theory of equilibrium selection in games. *MIT Press Books* 1
Heider F. 1958. The psychology of interpersonal relations. John Wiley & Sons
Heider F, Simmel M. 1944. An experimental study of apparent behavior. *The American journal of psychology* 57:243–259
Henrich J. 2004. Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization* 53:3–35
Henrich J. 2009. The evolution of costly displays, cooperation and religion: Credibility enhancing displays and their implications for cultural evolution. *Evolution and human behavior* 30:244–260
Herrmann B, Thoni C, Gachter S. 2008. Antisocial punishment across societies. *Science* 319:1362–1367
Hitchcock C, Knobe J. 2009. Cause and norm. *The Journal of Philosophy* 106:587–612
Ho MK, Abel D, Correa CG, Littman ML, Cohen JD, Griffiths TL. 2022a. People construct simplified mental representations to plan. *Nature* 606:129–136
Ho MK, Cushman F, Littman ML, Austerweil JL. 2019. People teach with rewards and punishments as communication, not reinforcements. *Journal of Experimental Psychology: General* 148:520
Ho MK, Cushman F, Littman ML, Austerweil JL. 2021. Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*
Ho MK, MacGlashan J, Littman ML, Cushman F. 2017. Social is special: A normative framework

for teaching with and learning from evaluative feedback. *Cognition* 167:91–106

Ho MK, Saxe R, Cushman F. 2022b. Planning with theory of mind. *Trends in Cognitive Sciences*

Hoffman M, Yoeli E, Nowak MA. 2015. Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences* 112:1727–1732

Holyoak KJ, Simon D. 1999. Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General* 128:3

Hornsby AN, Love BC. 2020. How decisions and the desire for coherency shape subjective preferences over time. *Cognition* 200:104244

Hsu M, Anen C, Quartz SR. 2008. The right and the good: distributive justice and neural encoding of equity and efficiency. *science* 320:1092–1095

Hutcherson CA, Bushong B, Rangel A. 2015. A neurocomputational model of altruistic choice and its implications. *Neuron* 87:451–462

Icard TF, Kominsky JF, Knobe J. 2017. Normality and actual causal strength. *Cognition* 161:80–93

Izuma K, Adolphs R. 2013. Social manipulation of preference in the human brain. *Neuron* 78:563–573

Izuma K, Saito DN, Sadato N. 2008. Processing of social and monetary rewards in the human striatum. *Neuron* 58:284–294

Izuma K, Saito DN, Sadato N. 2010. Processing of the incentive for social approval in the ventral striatum during charitable donation. *Journal of cognitive neuroscience* 22:621–631

Jara-Ettinger J. 2019. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences* 29:105–110

Jara-Ettinger J, Gweon H, Schulz LE, Tenenbaum JB. 2016. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences* 20:589–604

Jara-Ettinger J, Tenenbaum JB, Schulz LE. 2015. Not so innocent: Toddlers' inferences about costs and culpability. *Psychological science* 26:633–640

Jordan JJ, Hoffman M, Bloom P, Rand DG. 2016a. Third-party punishment as a costly signal of trustworthiness. *Nature* 530:473–476

Jordan JJ, Hoffman M, Nowak MA, Rand DG. 2016b. Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences* 113:8658–8663

Kahneman D. 2011. Thinking, fast and slow. macmillan

Kalkstein DA, Hook CJ, Hard BM, Walton GM. 2022. Social norms govern what behaviors come to mind—and what do not. *Journal of Personality and Social Psychology*

Kelley HH. 1973. The processes of causal attribution. *American psychologist* 28:107

Keramati M, Smittenaar P, Dolan RJ, Dayan P. 2016. Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences* 113:12868–12873

Kim M, Park B, Young L. 2020. The psychology of motivated versus rational impression updating. *Trends in Cognitive Sciences* 24:101–111

Kleiman-Weiner M, Gerstenberg T, Levine S, Tenenbaum JB. 2015. Inference of intention and permissibility in moral decision making., In *CogSci*

Kleiman-Weiner M, Ho MK, Austerweil JL, Littman ML, Tenenbaum JB. 2016. Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction, In *CogSci*

Kleiman-Weiner M, Saxe R, Tenenbaum JB. 2017. Learning a commonsense moral theory. *Cognition* 167:107–123

Knobe J, Nichols S. 2011. Free will and the bounds of the self

Koster R, Duzel E, Dolan RJ. 2015. Action and valence modulate choice and choice-induced preference change. *PLoS One* 10:e0119682

Kraft-Todd G, Kleiman-Weiner M, Young L. 2020. Differential discounting of virtue signaling: public virtue is perceived less favorably than private virtue for generosity but not impartiality

Krajbich I, Armel C, Rangel A. 2010. Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience* 13:1292–1298

Kunda Z. 1990. The case for motivated reasoning. *Psychological bulletin* 108:480
Lau T, Pouncy HT, Gershman SJ, Cikara M. 2018. Discovering social groups via latent structure learning. *Journal of Experimental Psychology: General* 147:1881
Le Pargneux A, Chater N, Zeitoun H. 2023. Contractualist reasoning influences moral judgment and decision making
Leong YC, Hughes BL, Wang Y, Zaki J. 2019. Neurocomputational mechanisms underlying motivated seeing. *Nature human behaviour* 3:962–973
Levine S, Kleiman-Weiner M, Schulz L, Tenenbaum J, Cushman F. 2020. The logic of universalization guides moral judgment. *Proceedings of the National Academy of Sciences* 117:26158–26169
Lewin K. 1936. Principles of topological psychology. McGraw-Hill
Lewis D. 2013. Counterfactuals. John Wiley & Sons
Lieder F, Griffiths TL. 2020. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences* 43:e1
Littman ML. 2001. Value-function reinforcement learning in markov games. *Cognitive systems research* 2:55–66
Liu S, Ullman TD, Tenenbaum JB, Spelke ES. 2017. Ten-month-old infants infer the value of goals from the costs of actions. *Science* 358:1038–1041
Lockwood PL, Klein-Flügge MC, Abdurahman A, Crockett MJ. 2020. Model-free decision making is prioritized when learning to avoid harming others. *Proceedings of the National Academy of Sciences* 117:27719–27730
Lombrozo T. 2010. Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive psychology* 61:303–332
Maier M, Cheung V, Bartoš F, Lieder F. 2023. Learning from consequences shapes reliance on moral rules vs. cost-benefit reasoning
Malle BF, Guglielmo S, Monroe AE. 2014. A theory of blame. *Psychological Inquiry* 25:147–186
Marr D. 1982. Vision: A computational investigation into the human representation and processing of visual information. MIT press
Martin JW, Cushman F. 2016. Why we forgive what can't be controlled. *Cognition* 147:133–143
Michotte A. 2017. The perception of causality. Routledge
Mikhail J. 2007. Universal moral grammar: Theory, evidence and the future. *Trends in cognitive sciences* 11:143–152
Milgram S, Bickman L, Berkowitz L. 1969. Note on the drawing power of crowds of different size. *Journal of personality and social psychology* 13:79
Morris A, Cushman F. 2018. A common framework for theories of norm compliance. *Social Philosophy and Policy* 35:101–127
Morris A, MacGlashan J, Littman ML, Cushman F. 2017. Evolution of flexibility and rigidity in retaliatory punishment. *Proceedings of the National Academy of Sciences* 114:10396–10401
Morris A, Phillips J, Huang K, Cushman F. 2021. Generating options and choosing between them depend on distinct forms of value representation. *Psychological science* 32:1731–1746
Morris A, Phillips JS, Icard T, Knobe J, Gerstenberg T, Cushman F. 2018. Causal judgments approximate the effectiveness of future interventions
Najar A, Bonnet E, Bahrami B, Palminteri S. 2020. The actions of others act as a pseudo-reward to drive imitation in the context of social reinforcement learning. *PLoS biology* 18:e3001028
Nakayama K, Shimojo S. 1992. Experiencing and perceiving visual surfaces. *Science* 257:1357–1363
Nash J. 1951. Non-cooperative games. *The Annals of Mathematics* 54:286–295
Nash JF. 1950. The bargaining problem. *Econometrica: Journal of the econometric society* :155–162
Newell A, Simon H. 1956. The logic theory machine–a complex information processing system. *IRE Transactions on information theory* 2:61–79
Nichols S. 2021. Rational rules: Towards a theory of moral learning. Oxford University Press
Nisbett RE. 2018. Culture of honor: The psychology of violence in the south. Routledge
Nowak MA. 2006. Five rules for the evolution of cooperation. *science* 314:1560–1563

O'Connor C. 2019. The origins of unfairness: Social categories and cultural evolution. Oxford University Press, USA

Olsson A, Phelps EA. 2007. Social learning of fear. *Nature neuroscience* 10:1095–1102

Ong DC, Zaki J, Goodman ND. 2019. Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science* 11:338–357

Ostrom E. 2000. Collective action and the evolution of social norms. *Journal of economic perspectives* 14:137–158

Park B, Kim M, Young L. 2021. An examination of accurate versus "biased" mentalizing in moral and economic decision-making. In *The Neural Basis of Mentalizing*. Springer, 537–554

Patil I, Zucchelli MM, Kool W, Campbell S, Fornasier F, et al. 2021. Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. *Journal of Personality and Social Psychology* 120:443

Pearl J. 2009. Causality. Cambridge university press

Phan KL, Sripada CS, Angstadt M, McCabe K. 2010. Reputation for reciprocity engages the brain reward center. *Proceedings of the National Academy of Sciences* 107:13099–13104

Phillips J, Cushman F. 2017. Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences* 114:4649–4654

Phillips J, Morris A, Cushman F. 2019. How we know what not to think. *Trends in cognitive sciences* 23:1026–1040

Pinker S. 2010. The cognitive niche: Coevolution of intelligence, sociality, and language. *Proceedings of the National Academy of Sciences* 107:8993–8999

Pizarro D, Uhlmann E, Salovey P. 2003. Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological science* 14:267–272

Quillien T, German TC. 2021. A simple definition of 'intentionally'. *Cognition* 214:104806

Quillien T, Lucas CG. 2022. Counterfactuals and the logic of causal selection

Rabin M. 1993. Incorporating fairness into game theory and economics. *The American economic review* :1281–1302

Rabinowitz N, Perbet F, Song F, Zhang C, Eslami SA, Botvinick M. 2018. Machine theory of mind, In *International conference on machine learning*. PMLR

Rand DG, Greene JD, Nowak MA. 2012. Spontaneous giving and calculated greed. *Nature* 489:427–430

Rand DG, Peysakhovich A, Kraft-Todd GT, Newman GE, Wurzbacher O, et al. 2014. Social heuristics shape intuitive cooperation. *Nature communications* 5:3677

Rangel A, Camerer C, Montague PR. 2008. A framework for studying the neurobiology of value-based decision making. *Nature reviews neuroscience* 9:545–556

Ratcliff R, Smith PL. 2004. A comparison of sequential sampling models for two-choice reaction time. *Psychological review* 111:333

Redish AD. 2022. Changing how we choose: The new science of morality. MIT Press

Rescorla RA. 1972. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical conditioning, Current research and theory* 2:64–69

Rilling JK, Sanfey AG. 2011. The neuroscience of social decision-making. *Annual review of psychology* 62:23–48

Roberts G. 1998. Competitive altruism: from reciprocity to the handicap principle. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 265:427–431

Ross L. 1977. The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology*, vol. 10. Elsevier, 173–220

Ruff CC, Fehr E. 2014. The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience* 15:549–562

Ruiz-Serra J, Harré MS. 2023. Inverse reinforcement learning as the algorithmic basis for theory of mind: Current methods and open problems. *Algorithms* 16:68

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD. 2003. The neural basis of economic

decision-making in the ultimatum game. *Science* 300:1755–1758
Saxe R, Houlihan SD. 2017. Formalizing emotion concepts within a bayesian model of theory of mind. *Current opinion in Psychology* 17:15–21
Schelling TC. 1960. The strategy of conflict: with a new preface by the author. Harvard university press
Schelling TC. 1971. Dynamic models of segregation. *Journal of mathematical sociology* 1:143–186
Sell A, Sznycer D, Al-Shawaf L, Lim J, Krauss A, et al. 2017. The grammar of anger: Mapping the computational architecture of a recalibrational emotion. *Cognition* 168:110–128
Sell AN. 2011. The recalibrational theory and violent anger. *Aggression and violent behavior* 16:381–389
Shafto P, Goodman ND, Griffiths TL. 2014. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology* 71:55–89
Sharot T, De Martino B, Dolan RJ. 2009. How choice reveals and shapes expected hedonic outcome. *Journal of Neuroscience* 29:3760–3765
Shin YS, Niv Y. 2021. Biased evaluations emerge from inferring hidden causes. *Nature human behaviour* 5:1180–1189
Shultz TR, Lepper MR. 1999. Computer simulation of cognitive dissonance reduction.
Shum M, Kleiman-Weiner M, Littman ML, Tenenbaum JB. 2019. Theory of minds: Understanding behavior in groups through inverse planning, In *Proceedings of the AAAI conference on artificial intelligence*, vol. 33
Siegel JZ, Mathys C, Rutledge RB, Crockett MJ. 2018. Beliefs about bad people are volatile. *Nature human behaviour* 2:750–756
Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, et al. 2017. Mastering the game of go without human knowledge. *nature* 550:354–359
Simon HA. 1956. Rational choice and the structure of the environment. *Psychological review* 63:129
Sloman SA. 1996. The empirical case for two systems of reasoning. *Psychological bulletin* 119:3
Smith JM. 1991. Honest signalling: the philip sidney game. *Animal Behaviour*
Smith JM. 1994. Must reliable signals always be costly? *Animal behaviour* 47:1115–1120
Smith JM, Price GR. 1973. The logic of animal conflict. *Nature* 246:15–18
Son JY, Bhandari A, FeldmanHall O. 2019. Crowdsourcing punishment: Individuals reference group preferences to inform their own punitive decisions. *Scientific reports* 9:1–15
Sosa FA, Ullman T, Tenenbaum JB, Gershman SJ, Gerstenberg T. 2021. Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition* 217:104890
Sullivan N, Hutcherson C, Harris A, Rangel A. 2015. Dietary self-control is related to the speed with which attributes of healthfulness and tastiness are processed. *Psychological science* 26:122–134
Sutton RS. 1988. Learning to predict by the methods of temporal differences. *Machine learning* 3:9–44
Sutton RS, Barto AG. 2018. Reinforcement learning: An introduction. MIT press
Swidler A. 1986. Culture in action: Symbols and strategies. *American sociological review* :273–286
Sznycer D. 2019. Forms and functions of the self-conscious emotions. *Trends in cognitive sciences* 23:143–157
Sznycer D, Tooby J, Cosmides L, Porat R, Shalvi S, Halperin E. 2016. Shame closely tracks the threat of devaluation by others, even across cultures. *Proceedings of the National Academy of Sciences* 113:2625–2630
Tamir DI, Thornton MA. 2018. Modeling the predictive social mind. *Trends in cognitive sciences* 22:201–212
Tappin BM, Pennycook G, Rand DG. 2020. Thinking clearly about causal inferences of politically motivated reasoning: Why paradigmatic study designs often undermine causal inference. *Current Opinion in Behavioral Sciences* 34:81–87
Tenenbaum JB, Griffiths TL, Kemp C. 2006. Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences* 10:309–318

Thompson B, Van Opheusden B, Sumers T, Griffiths T. 2022. Complex cognitive algorithms preserved by selective social learning in experimental populations. *Science* 376:95–98

Tinbergen N. 1963. On aims and methods of ethology. *Zeitschrift für tierpsychologie* 20:410–433

Todd PM, Gigerenzer G. 2000. Précis of simple heuristics that make us smart. *Behavioral and brain sciences* 23:727–741

Tomasello M. 2009. The cultural origins of human cognition. Harvard university press

Treadway MT, Buckholtz JW, Martin JW, Jan K, Asplund CL, et al. 2014. Corticolimbic gating of emotion-driven punishment. *Nature neuroscience* 17:1270–1275

Trivers RL. 1971. The evolution of reciprocal altruism. *The Quarterly review of biology* 46:35–57

Uhlmann EL, Pizarro DA, Diermeier D. 2015. A person-centered approach to moral judgment. *Perspectives on Psychological Science* 10:72–81

van Baar JM, Chang LJ, Sanfey AG. 2019. The computational and neural substrates of moral strategies in social decision-making. *Nature communications* 10:1483

Van den Bos W, van Dijk E, Westenberg M, Rombouts SA, Crone EA. 2009. What motivates repayment? neural correlates of reciprocity in the trust game. *Social cognitive and affective neuroscience* 4:294–304

Vasilyeva N, Blanchard T, Lombrozo T. 2018. Stable causal relationships are better causal relationships. *Cognitive Science* 42:1265–1296

Veblen T. 1889. The theory of the leisure class. George Allen Unwin, Ltd.

Vélez N, Chen A, , Burke T, Cushman F, Gershman S. 2023. Teachers recruit mentalizing regions to represent learners' beliefs. *Proceedings of the National Academy of Sciences* 38

Vélez N, Gweon H. 2019. Integrating incomplete information with imperfect advice. *Topics in cognitive science* 11:299–315

Vélez N, Gweon H. 2021. Learning from other minds: An optimistic critique of reinforcement learning models of social learning. *Current opinion in behavioral sciences* 38:110–115

Vinckier F, Rigoux L, Kurniawan IT, Hu C, Bourgeois-Gironde S, et al. 2019. Sour grapes and sweet victories: How actions shape preferences. *PLOS Computational Biology* 15:e1006499

Von Hippel W, Trivers R. 2011. The evolution and psychology of self-deception. *Behavioral and brain sciences* 34:1–16

Vul E, Pashler H. 2008. Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science* 19:645–647

Watkins CJ, Dayan P. 1992. Q-learning. *Machine learning* 8:279–292

Weiner B. 1995. Judgments of responsibility: A foundation for a theory of social conduct. guilford Press

Wu CM, Vélez N, Cushman FA. 2022. Representational exchange in human social learning. *The Drive for Knowledge: The Science of Human Information Seeking* :169

Young L, Cushman F, Hauser M, Saxe R. 2007. The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences* 104:8235–8240

Yu H, Siegel JZ, Clithero JA, Crockett MJ. 2021. How peer influence shapes value computation in moral decision-making. *Cognition* 211:104641

Yu H, Siegel JZ, Crockett MJ. 2019. Modeling morality in 3-d: Decision-making, judgment, and inference. *Topics in cognitive science* 11:409–432

Zahavi A. 1975. Mate selection—a selection for a handicap. *Journal of theoretical Biology* 53:205–214

Zaki J, Mitchell JP. 2011. Equitable decision making is associated with neural markers of intrinsic value. *Proceedings of the National Academy of Sciences* 108:19761–19766

Zaki J, Schirmer J, Mitchell JP. 2011. Social influence modulates the neural computation of value. *Psychological science* 22:894–900

Zhang Z, Wang S, Good M, Hristova S, Kayser AS, Hsu M. 2021. Retrieval-constrained valuation: Toward prediction of open-ended decisions. *Proceedings of the National Academy of Sciences* 118:e2022685118